Introduction to Big Data Management

Examination

Department of Computer Science
University of Helsinki

Instruction: Please answer all questions. There are five sections and the full mark is 50. No notes, computer, calculator, smartphone or any other material is allowed in the exam. Please write your name in the answer sheet.

Section A. For each question as follows, select one or multiple options. (1 mark per question)

A1. What are the three types of diverse data sources mentioned in this course? (Single choice)

A. Machine Data, Map Data, and Social Media

B. Sensor Data, Organizational Data, and Social Media

C. Information Networks, Map Data, and People

D. Machine Data, Organizational Data, and People

A2. What is the workflow for working with big data? (Single choice)

A. Big Data -> Better Models -> Higher Precision

B. Extrapolation -> Understanding -> Reproducing

C. Theory -> Models -> Precise Advice

A3. Consider three documents: d1 (*Helsinki Finland University*); d2 (*Helsinki Finland Europe*); d3 (*Computer Science University*); and a query q (*Computer Science Finland*). Compare the similarities with cosine TF-IDF. No need to compute the exact value to answer this question. (Single choice)

A. sim(d1,q)> sim(d2,q)> sim(d3,q)
B. sim(d3,q)> sim(d2,q)> sim(d1,q)
C. sim(d3,q)> sim(d1,q)> sim(d2,q)
D. sim(d2,q)> sim(d3,q)> sim(d1,q)

A4. Of the following, which procedure best generalizes big data procedures such as (but not limited to) the map reduce process? (Single choice)

A. split ->shuffle and sort->map->reduce
B. split->do->merge
C. split->map->shuffle and sort->reduce
D. split->sort->merge

A5. Which are the appropriate application scenarios for a MapReduce program? (Select 2 choices)

A. Perform the matrix multiplication and other complicated operations.

B. Run machine learning algorithms with many iterations.

C. Compute the inverted indices.

D. Summarize the number of pages crawled per host.

A6. What is the lambda architecture as shown in this course? (Single choice)

A. A type of hybrid data processing architecture.

B. A type of swappable data processing layer.

C. An architecture that natively supports lambda calculus.

D. A type of architecture that only contains part of the data processing method.

A7. What are the three layers in Lambda architecture? (Single choice)

A. Streaming Layer, Batch Layer, Serving Layer

B. Batch Layer , Speed Layer , Serving Layer

C. Storage Layer, Speed Layer, Serving Layer

D. Batch Layer, Storage Layer, Serving Layer

A8. Consider the following transaction involving two bank accounts x and y.

read(x); x := x – 50; write(x); read(y); y := y + 50; write(y)

The constraint that the sum of the accounts x and y should remain constant is that of (Single choice)

A. Atomicity

B. Consistency

C. Isolation

D. Durability

A9.What are the correct explanations on CAP Theorem? (Select 2 choices)

A. During normal operation (lack of network partition), the CAP theorem still imposes constraints on the availability or consistency of data.
B. Consistency in CAP Theorem has the same meaning as Consistent in ACID property of database transaction.
C. There are systems that are available and partition tolerant but cannot guarantee consistency.
D. CAP says that, in case of a network partition (a rare occurrence) one needs to choose between availability and partition tolerance.

A10. What are the correct statements on Strong (Strict) Consistency and Eventual Consistency? (Select 3 choices)

A. In the strong consistency, all replicas must be in the same state for the next operation to occur on any value.
B. In the eventual consistency, all read operations always return the value from the last finalized write operation.
C. In the strong consistency, at any given point readers will see some written value, and there is the guarantee that any two readers will see the exact same write.
D. In the eventual consistency, all replicas will eventually have the latest update; it's just a matter of time when that will happen.

A11. Which are the correct statements on the functions of Mapper and Reducer ? (Select 2 choices)

A. Each Mapper can do something to each individual key-value pair.

B. Each Mapper can look at key-value pairs of other mappers.

C. Each Reducer can aggregate data.

D. Each Reduce can look at multiple values from other reducers.

A12. What is the advantage of multi-model database over polyglot persistence? (Select 1 choice)

A. Polyglot persistence is just a theoretical idea, no practical solution.

B. Multi-model databases use a single engine for holistic query optimization.

C. Most databases become multi-model databases.

D. Polyglot persistence can support only a single data type.

A13. Which are NOT typical scenarios to consider the deployment of Hadoop system? (Select 2 choices)

A. You see a large scale growth in amount of data.
B. You perform the simultaneous execution of many different functions on multiple nodes across the same or different data sets.
C. You want quick access to your old data which would otherwise go on tape drives for archival storage.
D. You perform frequent random data access in application

A14. Select the correct statements on semi-join for distributed databases? (Select 2 choices)

A. The results of semi-join contain the columns from all of joined tables.
B. All SQL join queries can be solved entirely using semi-joins.
C. The difference between a semi-join and a traditional join is that rows in one of joined tables will be returned at most once.
D. Semi-join does not give the final result still it is efficient in reducing communication cost than conventional join.

Section B. Use your own words to define and explain the following concepts.

B1. Define and explain the differences between: Data Warehouse, Data Mart, Data Lake. (3 marks)

B2. Define and explain the differences between Database management system (DBMS) and Big data management system (BDMS) (3 marks)

B3. Define and explain the differences between data integration, data fusion and data exchange. (3 marks)

Section C. Write SQL queries in the following cases:

C1. Consider the following schema (the data type of each attribute is given following the attribute names and the primary keys are underlined):

Suppliers (**sid**: integer, sname: string, address: string)

Parts (**pid**: integer, pname: string, color: string)

Catalog (**sid**: integer, **pid**: integer, cost: real)

The Catalog relation lists the prices charged for parts by Suppliers. Write the following queries in SQL:

i. Find the sids of suppliers who supply a red part or a green part. (2 marks)
ii. Find the sids of suppliers who charge more for some parts than the average cost of that part (averaged over all the suppliers who supply that part). (2 marks)

C2. Write the views in SQL on the following database schema: (the key is underlined)

EMPLOYEE
(SSN , First_Name,Last_Name, ,BDate,Address,Sex,Salary,Supervisor_SSN,Department_NO)

DEPARTMENT
(Department_NO, Department_Name, Manager_SSN,Manager_Start_Date)

PROJECT
(Project_NO , Project_Name, Location,Department_NO)

WORKS_ON
(Employee_SSN, Project_NO, Hours)

a) A view that has the project name, controlling department name, number of employees and total hours worked per week on the project for each project; (2 marks)
b) A view that has the project name, controlling department name, number of employees, and total hours per week on the project for each project with more than one employee working on it. (2 marks)

C3. Write an integrated global view for the following movies data sources with SQL. (3 marks)

3 movie sources

S1(title,dir,year,genre) from until 1980.

S2(title,dir) since 1970

S3(title, year, genre) all movies

Global table:

S (title,dir,year,genre)

Define the global view with SQL for table S such that S = S1 union (S2 join S3).

Section D.

Consider a MongoDB database with a single collection 'customers', which stores information about the customers of a bank. For each customer, the collection contains one document with the customer's name (as a string), address (as a string), number of transactions (as integer), total volume of transactions in euros (as a real number), and the year the customer joined the bank (as integer).

The following is an example of such a document.

```
{
'name': "Iris Huhtala",
'address': "Aleksanterikatu 29, Helsinki, Finland, 00100",
'transactionsnum': 26,
'transactionsvolume': 300027,
'year': 2018
}
```

a. The data scientist of the bank tries to find customers with suspicious activity. Towards this end, she intends to find the names of customers who joined the bank after 2016 and have more than 1 million euros in transactions volume. To do this, she writes the following query.

```
db.customers.find( { 'year': { '$gt': 2016 } }, { '_id': 0 })
```

i. What results will the above query return? (1 mark)
ii. Do the results contain the information that the data scientist intends to find? (1 mark)
iii. Modify the above query so that it returns exactly the information that the data scientist intends to find. (1 mark)

b. The data scientist realizes that she will have to submit many similar queries like the above, to find the names of customers who joined the bank after some year X and have a transactions volume that is higher than some amount Y of euros. Many such queries will be submitted for different values X and Y.

To speed-up the execution time of queries, she considers building one of these indexes:

I. Index on field '*name*' (in ascending order).
II. Compound index on field '*transactionsnum*' (in ascending order) followed by field 'year' (in ascending order).
III. Index on field '*transactionsvolume*' (in ascending order).

Which of these indexes will be (most) helpful in speeding up the queries? Explain your answer.(2 marks)

Section E. Write pseudo-codes for the following MapReduce program and answer the questions.

E1. A MapReduce program, referred to as a job, consists of code for mappers and reducers. Write the pseudo-code for the *Wordcount* MapReduce program. What does the shuffle and sort procedure do when a MapReduce job is run? (5 marks)

E2. Consider a two-dimensional space where both the x and y coordinates range from -1000000000 to +1000000000. You have one file with the location of foos, and another file with the location of bars. Each record in those files is a comma-separated *(x,y)* coordinate. For example, a couple of lines may look like (*145999, 888888880*), (*834478899, 5656*). Write MapReduce pseudo-codes for mappers and reducers to find all foos that are less than 1 unit distance from a bar. Distance is measured using the familiar Euclidean distance, $sqrt[(x1-x2)^2 + (y1-y2)^2]$. Although both foos and bars are relatively sparse in this 2D space, their respective files are too big to be stored in memory. (6 marks)